**Stuart Russell** is a professor of computer science at the University of California, Berkeley, and an expert on artificial intelligence.

COMMENTARY

# SHOULD WE FEAR SUPERSMART ROBOTS?

## If we're not careful, we could find ourselves at odds with determined, intelligent machines whose objectives conflict with our own

*By Stuart Russell*

* * * * * * * * * * * * * * * * * * * * * * * * * * * *

IT IS HARD TO ESCAPE THE NAGGING SUSPICION THAT CREATING MACHINES smarter than ourselves *might* be a problem. After all, if gorillas had accidentally created humans way back when, the now endangered primates probably would be wishing they had not done so. But *why,* specifically, is advanced artificial intelligence a problem?

Hollywood's theory that spontaneously evil machine consciousness will drive armies of killer robots is just silly. The real problem relates to the possibility that AI may become incredibly good at achieving something other than what we really want. In 1960 legendary mathematician Norbert Wiener, who founded the field of cybernetics, put it this way: "If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere…, we had better be quite sure that the purpose put into the machine is the purpose which we really desire."

A machine with a specific purpose has another property, one that we usually associate with living things: a wish to preserve its own existence. For the machine, this trait is not innate, nor is it something introduced by humans; it is a logical consequence of the simple fact that the machine cannot achieve its original purpose if it is dead. So if we send out a robot with the sole directive of fetching coffee, it will have a strong incentive to ensure success by disabling its own off switch or even exterminating anyone who might interfere with its mission. If we are not careful, then, we could face a kind of global chess match against very determined, superintelligent machines whose objectives conflict with our own, with the real world as the chessboard.

The prospect of entering into and losing such a match should concentrate the minds of computer scientists. Some researchers argue that we can seal the machines inside a kind of fire wall, using them to answer difficult questions but never allowing them to affect the real world. (Of course, this means giving up on superintelligent robots!) Unfortunately, that plan seems unlikely to work: we have yet to invent a fire wall that is secure against ordinary humans, let alone superintelligent machines.

Can we instead tackle Wiener's warning head-on? Can we de-

sign AI systems whose goals do not conflict with ours so that we are sure to be happy with the way they behave? This is far from easy—after all, stories with a genie and three wishes often end with a third wish to undo the first two—but I believe it is possible if we follow three core principles in designing intelligent systems:

*The machine's purpose must be to maximize the realization of human values.* In particular, it has no purpose of its own and no innate desire to protect itself.
*The machine must be initially uncertain about what those human values are.* This turns out to be crucial, and in a way it sidesteps Wiener's problem. The machine may learn more about human values as it goes along, of course, but it may never achieve complete certainty.
*The machine must be able to learn about human values by observing the choices that we humans make.*

The first two principles may seem counterintuitive, but together they avoid the problem of a robot having a strong incentive to disable its own off switch. The robot is sure it wants to maximize human values, but it also does not know exactly what those are. Now the robot actually *benefits* from being switched off because it understands that the human will press the off switch to prevent the robot from doing something counter to human values. Thus, the robot has a positive incentive to keep the off switch intact—and this incentive derives directly from its uncertainty about human values.

The third principle borrows from a subdiscipline of AI called inverse reinforcement learning (IRL), which is specifically concerned with learning the values of some entity—whether a human, canine or cockroach—by observing its behavior. By watching a typical human's morning routine, the robot learns about the value of coffee to humans. The field is in its infancy, but already some practical algorithms exist that demonstrate its potential in designing smart machines.

As IRL evolves, it must find ways to cope with the fact that humans are irrational, inconsistent, weak-willed and have limited computational powers, so their actions do not always reflect their values. Also, humans exhibit diverse sets of values, which means that robots must be sensitive to potential conflicts and trade-offs among people. And some humans are just plain evil and should be neither helped nor emulated.

Despite these difficulties, I believe it will be possible for machines to learn enough about human values that they will not pose a threat to our species. Besides directly observing human behavior, machines will be aided by having access to vast amounts of written and filmed information about people doing things (and others reacting). Designing algorithms that can understand this information is much easier than designing superintelligent machines. Also, there are strong economic incentives for robots—and their makers—to understand and acknowledge human values: if one poorly designed domestic robot cooks the cat for dinner, not realizing that its sentimental value outweighs its nutritional value, the domestic robot industry will be out of business.

Solving the safety problem well enough to move forward in AI seems to be feasible but not easy. There are probably decades to plan for the arrival of superintelligent machines. But the problem should not be dismissed out of hand, as it has been by some AI researchers. Some argue that humans and machines can coexist as long as they work in teams—yet that is not feasible unless machines share the goals of humans. Others say we can just "switch them off" as if superintelligent machines are too stupid to think of that possibility. Still others think that superintelligent AI will never happen. On September 11, 1933, renowned physicist Ernest Rutherford stated, with utter confidence, "Anyone who expects a source of power in the transformation of these atoms is talking moonshine." On September 12, 1933, physicist Leo Szilard invented the neutron-induced nuclear chain reaction. SA